Optimal "Anti-Bayesian" Parametric Pattern Classification Using Order Statistics Criteria

A. Thomas and B. John Oommen*

School of Computer Science, Carleton University, Ottawa, Canada : K1S 5B6

Abstract. The gold standard for a classifier is the condition of optimality attained by the Bayesian classifier. Within a Bayesian paradigm, if we are allowed to compare the testing sample with only *a single* point in the feature space from each class, the *optimal* Bayesian strategy would be to achieve this based on the (Mahalanobis) distance from the corresponding means. The reader should observe that, in this context, the mean, in one sense, is the most *central* point in the respective distribution. In this paper, we shall show that we can obtain optimal results by operating in a diametrically opposite way, i.e., a so-called "anti-Bayesian" manner. Indeed, we shall show the completely counter-intuitive result that by working with a *very few* (sometimes as small as two) points *distant* from the mean, one can obtain remarkable classification accuracies. Further, if these points are determined by the *Order Statistics* of the distributions, the accuracy of our method, referred to as Classification by Moments of Order Statistics (CMOS), attains the optimal Bayes' bound! This claim, which is totally counter-intuitive, has been proven for many uni-dimensional, and some multi-dimensional distributions within the exponential family, and the theoretical results have been verified by rigorous experimental testing. Apart from the fact that these results are quite fascinating and pioneering in their own right, they also give a theoretical foundation for the families of Border Identification (BI) algorithms reported in the literature.

Keywords: Classification using Order Statistics (OS), Moments of OS.

1 Introduction

Pattern classification is the process by which unknown feature vectors are categorized into groups or classes based on their features [1]. The age-old strategy for doing this is based on a Bayesian principle which aims to maximize the *a posteriori* probability. It is well known that when expressions for the latter are simplified, the classification criterion which attains the Bayesian optimal lower

* *Chancellor's Professor* ; *Fellow: IEEE* and *Fellow: IAPR*. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway. The work of this author was partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada. This paper was presented as a Keynote/Plenary talk at the conference.

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 1–13, 2012.

bound often reduces to testing the sample point using the corresponding distances/norms to the *means* or the "central points" of the distributions.

In this paper, we shall demonstrate that we can obtain optimal results by operating in a diametrically opposite way, i.e., a so-called "anti-Bayesian" manner. Indeed, we shall show the completely counter-intuitive result that by working with a *few* points *distant* from the mean, one can obtain remarkable classification accuracies. The number of points referred to can be as small as *two* in the uni-dimensional case. Further, if these points are determined by the *Order Statistics* of the distributions, the accuracy attains the optimal Bayes' bound! Thus, put in a nut-shell, we introduce here the theory of optimal pattern classification using Order Statistics of the features rather than the distributions of the features themselves. Thus, we propose a novel methodology referred to as Classification by Moments of Order Statistics (CMOS). It turns out, though, that this process is computationally not any more complex than working with the latter distributions.

1.1 Contributions of This Paper

The novel contributions of this paper are the following:

- We propose an "anti-Bayesian" paradigm for the classification of patterns within the parametric mode of computation, where the distance computations are not with regard to the "mean" but with regard to some samples "distant" from the mean. These points, which are sometimes as few as *two*, are the moments of OS of the distributions;
- We provide a theoretical framework for adequately responding to the question of why the border points are more informative for the task of classification;
- To justify these claims, we submit a formal analysis and the results of various experiments which have been performed for many distributions within the exponential family, and the results are clearly conclusive.

We conclude by mentioning that our results probably represent the state-of-the-art in BI!

2 Relevant Background Areas

2.1 Prototype Reduction Schemes and Border Identification Algorithms

If we fast-forward the clock by five decades since the initial formulation of Pattern Recognition (PR) as a research field, the informed reader will also be aware of the development of efficient classification methods in which the schemes achieve their task based on a *subset* of the training patterns. These are commonly referred to as "Prototype Reduction Schemes" (PRS)[2,3]. A PRS will be considered to be a generic method for reducing the number of training vectors, while

simultaneously attempting to guarantee that the classifier built on the reduced design set performs as well, or nearly as well, as the classifier built on the original design set [4]. Thus, instead of considering all the training patterns for the classification, a subset of the whole set is selected based on certain criteria. The learning (or training) is then performed on this reduced training set, which is also called the "Reference" set. This Reference set not only contains the patterns which are closer to the true discriminant's boundary, but also the patterns from the other regions of the space that can adequately represent the entire training set. Zillions of PRS [5] techniques have developed over the years, and it is clearly impossible to survey all of these here. These include the Condensed Nearest Neighbor (CNN) rule [6], the Reduced Nearest Neighbor (RNN) rule [7], the Prototypes for Nearest Neighbor (PNN) classifiers [8], the Selective Nearest Neighbor (SNN) rule [9], the Edited Nearest Neighbor (ENN) rule [10], Vector Quantization (VQ) etc. Comprehensive survey of the state-of-the-art in PRSs can be found in [2,11,3]. The formal algorithms are also found in [12].

Border Identification (BI) algorithms, which form a distinct subset of PRSs, work with a Reference set which only contains "border" points. A PRS would attempt to determine the relevant samples in both the classes which are capable of achieving near-optimal classification. As opposed to this, a BI algorithm uses *only* those samples which lie close to the *boundaries* of the two classes. Important developments in this area were proposed by Duch [13], Foody [14] and Li [15]. Duch developed algorithms to obtain the reference set based on a border analysis of *every* training pattern, and those algorithms attempt to *add* patterns which are closer to the class boundary, to the reference set. According to Foody's approach, the training set is divided into two sets - the first comprising of the set of border patterns, and the second being the set of non-border patterns. A border training set should contain patterns from different classes, but which are close together in the feature space and which are thus in the proximity of the true classification boundary. According to Li, the border patterns obtained by the traditional approaches are considered to be the "Near" borders, and using the latter, the "Far" borders are identified from the remaining data points. It turns out that the final border points computed in this manner are more accurate than the initially identified "Near" borders. The "Near" and the "Far" borders collectively constitute the so-called "Full" border set for the training data. A detailed survey of these methods can be found in [12,16].

2.2 Order Statistics

Let $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ be a univariate random sample of size n that follows a continuous distribution function Φ, where the probability density function (pdf) is $\varphi(\cdot)$. Let $\mathbf{x}_{1,n}$, $\mathbf{x}_{2,n}$, ..., $\mathbf{x}_{n,n}$ be the corresponding Order Statistics (OS). The r^{th} OS, $\mathbf{x}_{r,n}$, of the set is the r^{th} smallest value among the given random variables. The pdf of $\mathbf{y} = \mathbf{x}_{r,n}$ is given by:

$$f_{\mathbf{y}}(y) = \frac{n!}{(r-1)!(n-r)!} \{\Phi(y)\}^{r-1} \{1-\Phi(y)\}^{n-r} \varphi(y),$$

where $r = 1, 2, ..., n$. The reasoning for the above expression is straightforward. If the r^{th} OS appears at a location given by $\mathbf{y} = \mathbf{x}_{r,n}$, it implies that the $r-1$ smaller elements of the set are drawn independently from a Binomial distribution with a probability $\Phi(y)$, and the other $n-r$ samples are drawn using the probability $1-\Phi(y)$. The factorial terms result from the fact that the $(r-1)$ elements can be independently chosen from the set of n elements.

Although the distribution $f_{\mathbf{y}}(y)$ contains all the information resident in $\mathbf{y}$, the literature characterizes the OS in terms of quantities which are of paramount importance, namely its moments [17]. To better appreciate the results presented later in this paper, an understanding of the moments of the OS is necessary. This is briefly presented below.

Using the distribution $f_{\mathbf{y}}(y)$, one can see that the k^{th} moment of $\mathbf{x}_{r,n}$ can be formulated as:

$$E[\mathbf{x}_{r,n}^k] = \frac{n!}{(r-1)!(n-r)!} \int_{-\infty}^{+\infty} y^k \Phi(y)^{k-1}(x)(1-\Phi(y))^{n-r}\varphi(y)dy,$$

provided that both sides of the equality exist [18,19].

The fundamental theorem concerning the OS that we invoke is found in many papers [20,19,17]. The result is merely cited below inasmuch as the details of the proof are irrelevant and outside the scope of this study. The theorem can be summarized as follows.

Let $n \geq r \geq k+1 \geq 2$ be integers. Then, since Φ is a nondecreasing and right-continuous function from $\mathbb{R} \to \mathbb{R}$, $\Phi(\mathbf{x}_{r,n})$ is uniform in [0,1]. If we now take the k^{th} moment of $\Phi(\mathbf{x}_{r,n})$, it has the form [20]:

$$E[\Phi^k(\mathbf{x}_{r,n})] = \frac{B(r+k, n-r+1)}{B(r, n-r+1)} = \frac{n!\ (r+k-1)!}{(n+k)!\ (r-1)!}, \tag{1}$$

where $B(a,b)$ denotes the *Beta* function, and $B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$ since its parameters are integers.

The above fundamental result can also be used for characterization purposes as follows [20]. Let $n \geq r \geq k+1 \geq 2$ be integers, with Φ being nondecreasing and right-continuous. Let G be *any* nondecreasing and right-continuous function from $\mathbb{R} \to \mathbb{R}$ on the same support as Φ. The relation

$$E[G^k(\mathbf{x}_{r,n})] = \frac{n!\ (r+k-1)!}{(n+k)!\ (r-1)!} \tag{2}$$

holds if and only if $\forall x, \Phi(x) = G(x)$. In other words, $\Phi(\cdot)$ is the unique function that satisfies Eq. (2), implying that every distribution is characterized by the moments of its OS.

The implications of the above are the following:

1. If $n = 1$, implying that only a *single* sample is drawn from $\mathbf{x}$, from Eq. (1),

$$E[\Phi^1(\mathbf{x}_{1,1})] = \frac{1}{2}, \implies E[\mathbf{x}_{1,1}] = \Phi^{-1}\left(\frac{1}{2}\right). \tag{3}$$

Informally speaking, the first moment of the 1-order OS would be the value where the cumulative distribution Φ equals $\frac{1}{2}$, which is the Median($\mathbf{x}$).

2. If $n = 2$, implying that only *two* samples are drawn from $\mathbf{x}$, we can deduce from Eq. (1) that:

$$E[\Phi^1(\mathbf{x}_{1,2})] = \frac{1}{3}, \implies E[\mathbf{x}_{1,2}] = \Phi^{-1}\left(\frac{1}{3}\right), \text{and} \tag{4}$$

$$E[\Phi^1(\mathbf{x}_{2,2})] = \frac{2}{3}, \implies E[\mathbf{x}_{2,2}] = \Phi^{-1}\left(\frac{2}{3}\right). \tag{5}$$

Thus, from a computational perspective, the first moment of the first and second 2-order OS would be the values where the cumulative distribution Φ equal $\frac{1}{3}$ and $\frac{2}{3}$ respectively.

Although the analogous expressions can be derived for the higher order OS, for the rest of this paper we shall merely focus on the 2-order OS, and derive the consequences of using them in classification!

3 Optimal Bayesian Classification Using *Two* Order Statistics

3.1 The Generic Classifier

Having characterized the moments of the OS of arbitrary distributions, we shall now consider how they can be used to design a classifier.

Let us assume that we are dealing with the 2-class problem with classes ω_1 and ω_2, where their class-conditional densities are $f_1(x)$ and $f_2(x)$ respectively (i.e, their corresponding distributions are $F_1(x)$ and $F_2(x)$ respectively)[1]. Let ν_1 and ν_2 be the corresponding *medians* of the distributions. Then, classification based on ν_1 and ν_1 would be the strategy that classifies samples based on a *single* OS. We shall show the fairly straightforward result that for all symmetric distributions, the classification accuracy of this classifier attains the Bayes' accuracy.

This result is not too astonishing because the median is centrally located close to (if not exactly) on the mean. The result for higher order OS is actually far more intriguing because the higher order OS are not located centrally (close to the means), but rather distant from the means. Consequently, we shall show that for a large number of distributions, mostly from the exponential family, the classification based on *these* OS again attains the Bayes' bound.

We shall initiate this discussion by examining the Uniform distribution. The reason for this is that even though the distribution itself is rather trivial, the analysis will provide the reader with an insight into the mechanism by which the problem can be tackled, which can then be extended for other distributions.

[1] Throughout this section, we will assume that the *a priori* probabilities are equal. If they are unequal, the above densities must be weighted with the respective *a priori* probabilities.

3.2 The Uniform Distribution

The continuous Uniform distribution is characterized by a constant function $U(a,b)$, where a and b are the minimum and the maximum values that the random variable $\mathbf{x}$ can take. If the class conditional densities of ω_1 and ω_2 are uniformly distributed,

$$f_1(x) = \begin{cases} \frac{1}{b_1 - a_1} & \text{if } a_1 \leq x \leq b_1; \\ 0 & \text{if } x < a_1 \text{ or } x > b_1, \text{ and} \end{cases}$$

$$f_2(x) = \begin{cases} \frac{1}{b_2 - a_2} & \text{if } a_2 \leq x \leq b_2; \\ 0 & \text{if } x < a_2 \text{ or } x > b_2. \end{cases}$$

The reader should observe the following:

- If $a_2 > b_1$, the two distributions are non-overlapping, rendering the classification problem trivial.
- If $a_2 < b_1$, but $b_1 - a_1 \neq b_2 - a_2$, the optimal Bayesian classification is again dependent only on the heights of the distributions. In other words, if $b_2 - a_2 < b_1 - a_1$, the testing sample will be assigned to ω_2 whenever $x > a_2$. This criterion again is not related to the mean of the distributions at all, and is thus un-interesting to our current investigations.
- The meaningful scenario is when $b_1 - a_1$ is exactly equal to $b_2 - a_2$, and if $a_2 < b_1$. In this case, the heights of the two distributions are equal and the distributions are overlapping. This is really the interesting case, and corresponds to the scenario when the two distributions are identical. We shall analyze this in greater detail and demonstrate that the optimal Bayesian classification is also attained by using the OS.

Theoretical Analysis: Uniform Distribution - 2-OS. We shall now derive the formal properties of the classifier that utilizes the OS for the Uniform distribution.

Theorem 1. *For the 2-class problem in which the two class conditional distributions are Uniform and identical, CMOS, the classification using two OS, attains the optimal Bayes' bound.*

Proof. The proof of the result is done in two steps. We shall first show that when the two class conditional distributions are Uniform and identical, the optimal Bayesian classification is achieved by a comparison to the corresponding *means*. The equivalence of this to a comparison to the corresponding OS leads to the final result.

Without loss of generality let the class conditional distributions for ω_1 and ω_2 be $U(0,1)$ and $U(h,1+h)$, with means $\mu_1 = \frac{1}{2}$ and $\mu_2 = h + \frac{1}{2}$, respectively. In this case, the optimal Bayes' classifier assigns x to ω_1 whenever $x < h$, x to ω_2 whenever $x > 1$, and x to ω_1 and to ω_2 with equal probability when $h < x < 1$. Since:

$$\begin{aligned} D\left(x,\ \mu_1\right) < D\left(x,\ \mu_2\right) &\iff x - \frac{1}{2} < h + \frac{1}{2} - x \\ &\iff 2x < 1 + h \\ &\iff x < \frac{1+h}{2}, \end{aligned} \tag{6}$$

we see that the optimal Bayesian classifier assigns the sample based on the proximity to the corresponding mean, proving the first assertion.

We now consider the moments of the OS of the distributions. If $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$ are n independent univariate random variables that follow the Uniform distribution $U(0,1)$, by virtue of Eq.(1), the expected values of the first moment of the k-order OS can be seen to be $E[\mathbf{x}_{k,n}] = \frac{k}{n+1}$. Thus, for $U(0,1)$, $E[\mathbf{x}_{1,2}] = \frac{1}{3}$ and $E[\mathbf{x}_{2,2}] = \frac{2}{3}$. Similarly, for the distribution $U(h, 1+h)$, the expected values are $E[\mathbf{x}_{1,2}] = h + \frac{1}{3}$ and $E[\mathbf{x}_{2,2}] = h + \frac{2}{3}$.

The OS-based classification is thus as follows: Whenever a testing sample comes from these distributions, the CMOS will compare the testing sample with $E[\mathbf{x}_{2,2}]$ of the first distribution, i.e., $\frac{2}{3}$, and with $E[\mathbf{x}_{1,2}]$ of the second distribution, i.e., $h + \frac{1}{3}$, and the sample will be labeled with respect to the class which minimizes the corresponding quantity. Observe that for the above rule to work, we must enforce the ordering of the OS of the two distributions, and this requires that $\frac{2}{3} < h + \frac{1}{3} \implies h > \frac{1}{3}$.

In order to prove that for $h > \frac{1}{3}$ the OS-based classification is identical to the mean-based classification, we have to prove that $D(x, \mu_1) < D(x, \mu_2) \implies D(x, O_1) < D(x, O_2)$, where O_1 is $E[\mathbf{x}_{2,2}]$ of the first distribution and O_2 is $E[\mathbf{x}_{1,2}]$ of the second distribution. By virtue of Eq.(6),

$$D(x, \mu_1) < D(x, \mu_2) \iff x < \frac{h+1}{2}. \tag{7}$$

Similarly,

$$\begin{aligned} D(x, O_1) < D(x, O_2) &\iff D\left(x,\ \frac{2}{3}\right) \ < \ D\left(x,\ h + \frac{1}{3}\right) \\ &\iff x - \frac{2}{3} \ < \ h + \frac{1}{3} - x \\ &\iff x \ < \ \frac{h+1}{2}. \end{aligned} \tag{8}$$

The result follows by observing that (7) and (8) are identical comparisons.

For the analogous result for the case when $h < \frac{1}{3}$, the CMOS will compare the testing sample with $E[\mathbf{x}_{1,2}]$ of the first distribution, i.e., $\frac{1}{3}$, and with $E[\mathbf{x}_{2,2}]$ of the second distribution, i.e., $h + \frac{2}{3}$. Again, the sample will be labeled with respect to the class which minimizes the corresponding quantity. The proofs of the equivalence of this to the Bayesian decision follows along the same lines as the case when $h > \frac{1}{3}$, and is omitted to avoid repetition.

Hence the theorem! □

Experimental Results: Uniform Distribution - 2-OS. The CMOS method explained in Section 3.2 has been rigorously tested for various uniform distributions with 2-OS. In the interest of brevity, a few typical results are given below.

For each of the experiments, we generated 1,000 points for the classes ω_1 and ω_2 characterized by $U(0,1)$ and $U(h,1+h)$ respectively. We then invoked a classification procedure by utilizing the Bayesian and the CMOS strategies. In every case, CMOS was compared with the Bayesian classifier for different values of h, as tabulated in Table 1. The results in Table 1 were obtained by executing each algorithm 50 times using a 10-fold cross-validation scheme.

Table 1. Classification of Uniformly distributed classes by the CMOS 2-OS method for different values of h

$\boldsymbol{h}$	0.95	0.90	0.85	0.80	0.75	0.70
Bayesian	97.58	95.1	92.42	90.23	87.82	85.4
CMOS	97.58	95.1	92.42	90.23	87.82	85.4

Observe that in every case, the accuracy of CMOS attained the Bayes' bound.

By way of example, we see that CMOS should obtain the Bayesian bound for the distributions $U(0,1)$ and $U(0.8,1.8)$ whenever $n < \frac{1+0.8}{1-0.8} = 9$. In this case, the expected values of the moments are $\frac{1}{10}$ and $\frac{9}{10}$ respectively. These results justify the claim of Theorem 1.

Theoretical Analysis: Uniform Distribution - k-OS. We have seen from Theorem 1 that the moments of the 2-OS are sufficient for the classification to attain a Bayes' bound. We shall now consider the scenario when we utilize other k-OS. The formal result pertaining to this is given in Theorem 2.

Theorem 2. *For the 2-class problem in which the two class conditional distributions are Uniform and identical as U(0,1) and U(h, 1+h), optimal Bayesian classification can be achieved by using symmetric pairs of the n-OS, i.e., the $n-k$ OS for ω_1 and the k OS for ω_2 if and only if $k > \frac{(n+1)(1-h)}{2}$.*

Proof. We know that for the uniform distribution $U(0,1)$, the expected values of the first moment of the k-order OS have the form $E[\mathbf{x}_{k,n}] = \frac{k}{n+1}$. Our claim is based on the classification in which we can choose any of the symmetric pairs of the n-OS, i.e., the $n-k$ OS for ω_1 and the k OS for ω_2, whose expected values are $\frac{n-k+1}{n+1}$ and $h + \frac{k}{n+1}$ respectively.

Consider the case when $h > 1 - \frac{2k}{n+1}$, the relevance of which will be argued presently. Whenever a testing sample comes, it will be compared with the corresponding k-OS symmetric pairs of the expected values of the n-OS, and the sample will be labeled with respect to the class that minimizes the distance.

Observe that for the above rule to work, we must again enforce the ordering of the OS of the two distributions, and this requires that:

$$\frac{n-k+1}{n+1} < h + \frac{k}{n+1} \implies k > \frac{(n+1)(1-h)}{2}. \tag{9}$$

Eq.(9) can be seen to be:

$$k > \frac{(n+1)(1-h)}{2} \implies h > 1 - \frac{2k}{n+1}, \tag{10}$$

which justifies the case under consideration. As we have already proved that the Bayesian bound can be achieved by a comparison to the corresponding means (in Eq.(6)), which in turn simplifies to $x \sim \omega_1 \iff x < \frac{h+1}{2}$, we need to show that to obtain optimal accuracy using these symmetric $n-k$ and k OS, $D(x, O_1) < D(x, O_2) \iff x < \frac{h+1}{2}$. Indeed, the OS-based classification also attains the Bayesian bound because:

$$\begin{aligned} D(x,O_1) < D(x,O_2) &\iff D\left(x,\ \frac{n-k+1}{n+1}\right) < D\left(x,\ h+\frac{k}{n+1}\right) \\ &\iff x - \frac{n-k+1}{n+1} < h + \frac{k}{n+1} - x \\ &\iff x < \frac{h+1}{2}. \end{aligned} \tag{11}$$

For the symmetric argument when $h < 1 - \frac{2k}{n+1}$, the CMOS will compare the testing sample with $E[\mathbf{x}_{k,n}]$ of the first distribution and $E[\mathbf{x}_{n-k,n}]$ of the second distribution and the classification is obtained based on the class that minimizes *this* distance. The details of the proof are analogous and omitted. Hence the theorem! □

An alternate methodology to visualize the theorem and its consequences is given in [12,16], and is omitted here in the interest of space.

Experimental Results: Uniform Distribution - *k*-OS. The CMOS method has also been tested for the Uniform distribution for other k OS. In the interest of brevity, we merely cite one example where the distributions for ω_1 and ω_2 were characterized by $U(0,1)$ and $U(0.8,1.8)$ respectively. For each of the experiments, we generated 1,000 points for each class, and the testing samples were classified based on the selected *symmetric* pairs for values k and $n-k$ respectively. The results are displayed in Table 2.

To clarify the table, consider the row given by Trial No. 6 in which the 7-OS were invoked for the classification. Observe that the k-OS are now given by $\frac{n-k+1}{n+1}$ and $\frac{k}{n+1}$ respectively. In this case, the possible symmetric OS pairs could be $\langle 1,6\rangle$, $\langle 2,5\rangle$, and $\langle 3,4\rangle$ respectively. In every single case, the accuracy attained the Bayes' bound, as indicated by the results in the table.

The consequence of violating the condition imposed by Theorem 2 can be seen from the results given in the row denoted by Trial No. 9. In this case, the testing

Table 2. Results of the classification obtained by using the symmetric pairs of the OS for different values of n. The value of h was set to be 0.8. Note that in every case, the accuracy attained the Bayes' value whenever the conditions stated in Theorem 2 were satisfied.

Trail No.	**Order(n)**	**Moments**	OS_1	OS_2	**CMOS**	**Pass/Fail**
1	Two	$\{\frac{i}{3} \mid 1 \leq i \leq 2\}$	$\frac{2}{3}$	$h + \frac{1}{3}$	90.23	Passed
2	Three	$\{\frac{i}{4} \mid 1 \leq i \leq 3\}$	$\frac{3}{4}$	$h + \frac{1}{4}$	90.23	Passed
3	Four	$\{\frac{i}{5} \mid 1 \leq i \leq 4\}$	$\frac{4}{5}$	h + $\frac{1}{5}$	90.23	Passed
4	Five	$\{\frac{i}{6} \mid 1 \leq i \leq 5\}$	$\frac{4}{6}$	$h + \frac{2}{6}$	90.23	Passed
5	Six	$\{\frac{i}{7} \mid 1 \leq i \leq 6\}$	$\frac{4}{7}$	$h + \frac{2}{7}$	90.23	Passed
6	Seven	$\{\frac{i}{8} \mid 1 \leq i \leq 7\}$	$\frac{5}{8}$	$h + \frac{3}{8}$	90.23	Passed
7	Eight	$\{\frac{i}{9} \mid 1 \leq i \leq 8\}$	$\frac{6}{9}$	$h + \frac{3}{9}$	90.23	Passed
8	Nine	$\{\frac{i}{10} \mid 1 \leq i \leq 9\}$	$\frac{7}{10}$	$h + \frac{3}{10}$	90.23	Passed
9	Ten	$\{\frac{i}{11} \mid 1 \leq i \leq 10\}$	$\frac{10}{11}$	$h + \frac{1}{11}$	9.77	Failed
10	Ten	$\{\frac{i}{11} \mid 1 \leq i \leq 10\}$	$\frac{9}{11}$	$h + \frac{2}{11}$	90.23	Passed
11	Ten	$\{\frac{i}{11} \mid 1 \leq i \leq 10\}$	$\frac{7}{11}$	$h + \frac{4}{11}$	90.23	Passed
12	Ten	$\{\frac{i}{11} \mid 1 \leq i \leq 10\}$	$\frac{6}{11}$	$h + \frac{5}{11}$	90.23	Passed

attained the Bayes accuracy for the symmetric OS pairs $\langle 2, 9\rangle$, $\langle 3, 8\rangle$, $\langle 4, 7\rangle$ and $\langle 5, 6\rangle$ respectively. However, the classifier "failed" for the specific 10-OS, when the OS used were $\frac{10}{11}$ and $h + \frac{1}{11}$, as these values did not satisfy the condition $h > 1 - \frac{2k}{n+1}$. Observe that if $h < 1 - \frac{2k}{n+1}$, the symmetric pairs should be reversed, i.e., $\frac{k}{n+1}$ for the first distribution, and $h + \frac{n-k+1}{n+1}$ for the second distribution, to obtain the optimal Bayesian bound. The astonishing facet of this result is that one obtains the Bayes accuracy even though the classification requires only *two* points distant from the mean, justifying the rationale for BI schemes, and yet operating in an anti-Bayesian manner!

Remark: We believe that the CMOS, the classification by the moments of Order Statistics, is also true for multi-dimensional distributions. For a *prima facie* case, we consider two (overlapping) 2-dimensional uniform distributions U_1 and U_2 in which both the features are in $[0, 1]^2$ and $[h, 1 + h]^2$ respectively. Consequently, we can see that the overlapping region of the distributions forms a square. In this case, it is easy to verify that the Bayesian classifier is the diagonal that passes through the intersection points of the distributions. For the classification based on the moments of the 2-OS, because the features are independent for both dimensions, we can show that this is equivalent to utilizing the OS at position $\frac{2}{3}$ of the first distribution for both dimensions, and the OS at the position $h + \frac{1}{3}$ of the second distribution for both dimensions.

Table 3. Classification of Uniformly distributed 2-dimensional classes by the CMOS 2-OS method for different values of h. In the last two cases, the OS points of interest are reversed as explained in Section 3.2.

$\boldsymbol{h}$	0.95	0.90	0.85	0.80	0.75	0.70	0.65	0.60
Bayesian	99.845	99.505	98.875	98.045	97.15	95.555	94.14	91.82
CMOS	99.845	99.505	98.875	98.045	97.15	95.555	94.14	91.82

The CMOS method for 2-dimensional uniform distributions U_1 (in $[0,1]$ in both dimensions) and U_2 (in $[h, 1+h]$ in both dimensions) has been rigorously tested, and the results are given in Table 3. A formal proof for the case when the second class is distributed in $[h_1, 1+h_1] \times [h_2, 1+h_2]$, and for multi-dimensional features is currently being devised. It will appear in a forthcoming paper.

3.3 The Laplace (or Doubly-Exponential) Distribution

The *Laplace distribution* is a continuous uni-dimensional pdf named after Pierre-Simon Laplace. It is sometimes called the *doubly exponential distribution*, because it can be perceived as being a combination of two exponential distributions, with an additional location parameter, spliced together back-to-back.

If the class conditional densities of ω_1 and ω_2 are doubly exponentially distributed,

$$f_1(x) = \frac{\lambda_1}{2} e^{-\lambda_1 |x - c_1|}, \quad -\infty < x < \infty, \text{and}$$

$$f_2(x) = \frac{\lambda_2}{2} e^{-\lambda_2 |x - c_2|}, \quad -\infty < x < \infty,$$

where c_1 and c_2 are the respective means of the distributions. By elementary integration and straightforward algebraic simplifications, the variances of the distributions can be seen to be $\frac{2}{\lambda_1^2}$ and $\frac{2}{\lambda_2^2}$ respectively.

If $\lambda_1 \neq \lambda_2$, the samples can be classified based on the heights of the distributions and their point of intersection. The formal results for the general case are a little more complex. However, to prove the analogous results of Theorem 1 for the Uniform distribution, we shall first consider the case when $\lambda_1 = \lambda_2$. In this scenario, the reader should observe the following:

- Because the distributions have the equal height, i.e. $\lambda_1 = \lambda_2$, the testing sample $\mathbf{x}$ will obviously be assigned to ω_1 if it is less than c_1 and be assigned to ω_2 if it is greater than c_2.
- Further, the crucial case is when $c_1 < \mathbf{x} < c_2$. In this regard, we shall analyze the CMOS classifier and prove that it attains the Bayes' bound even when one uses as few as *only* 2 OSs.

Theoretical Analysis: Doubly-Exponential Distribution. By virtue of Eq. (4) and (5), the expected values of the first moments of the two OS can be obtained by determining the points where the cumulative distribution function attains the values $\frac{1}{3}$ and $\frac{2}{3}$. Let u_1 be the point for the percentile $\frac{2}{3}$ of the first distribution, and u_2 be the point for the percentile $\frac{1}{3}$ of the second distribution. These points can be obtained as $u_1 = c_1 - \frac{1}{\lambda_1}\ log\left(\frac{2}{3}\right)$ and $u_2 = c_2 + \frac{1}{\lambda_2}\ log\left(\frac{2}{3}\right)$. With these points at hand, we can demonstrate that, for doubly exponential distributions, the classification based on the expected values of the moments of the 2-OS, CMOS, attains the Bayesian bound, and the proof can be seen in [21].

A similar argument can be raised for the classification based on the k-OS. For the 2-class problem in which the two class conditional distributions are Doubly Exponential and identical, the optimal Bayesian classification can be achieved by using symmetric pairs of the n-OS, i.e., the $n-k$ OS for ω_1 and the k OS for ω_2 if and only if $log\left(\frac{2k}{n+1}\right) > \frac{c_1-c_2}{2}$, and this claim is also proved in [12,16].

Analogous results for the uni-dimensional Gaussian distribution are also available, but omitted here, in the interest of brevity. They can be found in [12,16].

4 Conclusions

In this paper, we have shown that the optimal Bayes' bound can be obtained by an "anti-Bayesian" approach named CMOS, Classification by Moments of Order Statistics. We have proved that the classification can be attained by working with a *very few* (sometimes as small as two) points *distant* from the mean. Further, if these points are determined by the *Order Statistics* of the distributions, the optimal Bayes' bound can be attained. The claim has been proved for many uni-dimensional distributions within the exponential family. The corresponding results for some multi-dimensional distributions have been alluded to, and the theoretical results have been verified by rigorous experimental testing. Apart from the fact that these results are quite fascinating and pioneering in their own right, they also give a theoretical foundation for the families of Border Identification (BI) algorithms reported in the literature.

References

1. Duda, R.O., Hart, P.: Pattern Classification and Scene Analysis. A Wiley Interscience Publication (2000)
2. Garcia, S., Derrac, J., Cano, J.R., Herrera, F.: Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. IEEE Transactions on Pattern Analysis and Machine Intelligence
3. Triguero, I., Derrac, J., Garcia, S., Herrera, F.: A Taxonomy and Experimental Study on Prototype Generation for Nearest Neighbor Classification. IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews
4. Kim, S., Oommen, B.J.: On Using Prototype Reduction Schemes and Classifier Fusion Strategies to Optimize Kernel-Based Nonlinear Subspace Methods. IEEE Transactions on Pattern Analysis and machine Intelligence 27, 455–460 (2005)

5. Kuncheva, L.I., Bezdek, J.C., Duin, R.P.W.: Decision Templates for Multiple Classifier Fusion: An Experimental Comparison. Pattern Recognition - The Journal of the Pattern Recognition Society 34, 299–314 (2001)
6. Hart, P.E.: The Condensed Nearest Neighbor Rule. IEEE Transactions on Information Theory 14, 515–516 (1968)
7. Gates, G.W.: The Reduced Nearest Neighbor Rule. IEEE Transactions on Information Theory 18, 431–433 (1972)
8. Chang, C.L.: Finding Prototypes for Nearest Neighbor Classifiers. IEEE Transactions on Computing 23, 1179–1184 (1974)
9. Ritter, G.L., Woodruff, H.B., Lowry, S.R., Isenhour, T.L.: An Algorithm for a Selective Nearest Neighbor Rule. IEEE Transactions on Information Theory 21, 665–669 (1975)
10. Devijver, P.A., Kittler, J.: On the Edited Nearest Neighbor Rule. In: Fifth International Conference on Pattern Recognition, pp. 72–80 (December 1980)
11. http://sci2s.ugr.es/pr/
12. Thomas, A.: Pattern Classification using Novel Order Statistics and Border Identification Methods. PhD thesis, School of Computer Science, Carleton University (to be submitted, 2013)
13. Duch, W.: Similarity based methods: a general framework for Classification, Approximation and Association. Control and Cybernetics 29(4), 937–968 (2000)
14. Foody, G.M.: Issues in Training Set Selection and Refinement for Classification by a Feedforward Neural Network. In: Proceedings of IEEE International Geoscience and Remote Sensing Symposium, pp. 409–411 (1998)
15. Li, G., Japkowicz, N., Stocki, T.J., Ungar, R.K.: Full Border Identification for Reduction of Training Sets. In: Proceedings of the Canadian Society for Computational Studies of Intelligence, 21st Conference on Advances in Artificial Intelligence, pp. 203–215 (2008)
16. Thomas, A., Oommen, B.J.: The Foundational Theory of Optimal "Anti-Bayesian" Parametric Pattern Classification Using Order Statistics Criteria (to be submitted, 2012)
17. Too, Y., Lin, G.D.: Characterizations of Uniform and Exponential Distributions. Academia Sinica 7(5), 357–359 (1989)
18. Ahsanullah, M., Nevzorov, V.B.: Order Statistics: Examples and Exercises. Nova Science Publishers, Inc. (2005)
19. Morris, K.W., Szynal, D.: A goodness-of-fit for the Uniform Distribution based on a Characterization. Journal of Mathematical Science 106, 2719–2724 (2001)
20. Lin, G.D.: Characterizations of Continuous Distributions via Expected values of two functions of Order Statistics. Sankhya: The Indian Journal of Statistics 52, 84–90 (1990)
21. Thomas, A., Oommen, B.J.: Optimal "Anti-Bayesian" Parametric Pattern Classification for the Exponential Family Using Order Statistics Criteria (to be submitted, 2012)